Using AI Uncertainty Quantification to Improve Human Decision-Making

Laura R. Marusich [1] **Jonathan Z. Bakdash** [2] **Yan Zhou** [2] **Murat Kantarcioglu** [2]

Abstract

AI Uncertainty Quantification (UQ) has the potential to improve human decision-making beyond AI predictions alone by providing additional probabilistic information to users. The majority of past research on AI and human decision-making has concentrated on model explainability and interpretability, with little focus on understanding the potential impact of UQ on human decision-making. We evaluated the impact on human decision-making for instance-level UQ, calibrated using a strict scoring rule, in two online behavioral experiments. In the first experiment, our results showed that UQ was beneficial for decision-making performance compared to only AI predictions. In the second experiment, we found UQ had generalizable benefits for decision-making across a variety of representations for probabilistic information. These results indicate that implementing high quality, instance-level UQ for AI may improve decision-making with real systems compared to AI predictions alone.

1. Introduction

Using AI to improve human decision-making requires effective human-AI interaction. Recent work on human-AI interaction guidelines focuses on explainability and interpretability (Amershi et al., 2019), which may improve subjective human ratings of trust in and usability of AI. However, a quantitative synthesis of studies found that explanations may not generally improve decision accuracy beyond AI prediction alone (Schemmer et al., 2022) in many application domains. One less-explored possibility for promoting effective human-AI interaction is AI Uncertainty Quantification (UQ) for predictions. AI UQ is posited to be key for human decision-making (Abdar et al., 2021b; Jalaian et al., 2019). However, there is conflicting evidence in the existing literature as to whether presenting AI UQ for predictions can improve human decision-making accuracy, and how to best communicate this uncertainty information (Lai et al., 2021). These conflicting results may be due in part to "a lack of discussion on the reliability of uncertainty estimates, sometimes referred to as calibration" (Lai et al., 2021, p. 15).

In order to resolve these questions, we use well-calibrated, instance-level AI Uncertainty Quantification (UQ) evaluated using a strict scoring rule (Gneiting & Raftery, 2007) using the ground truth for class labels [1]. We evaluate the impact of this AI UQ in two pre-registered, large sample size, online behavioral experiments assessing human decision-making. Decision-making is measured objectively using response accuracy and confidence calibration with accuracy. We found that providing high-quality AI UQ meaningfully improves decision-accuracy and confidence calibration over an AI prediction alone. Additionally, the benefits of this AI UQ appear to be generalizable – decision-making was similar for AI UQ presented with different visualizations and types of information. Our results indicate well-calibrated AI UQ is beneficial for decision-making.

The paper is structured as follows. In section 2, we provide the background information on uncertainty and human decision-making and an overview of existing techniques for AI UQ. Section 3 describes our UQ technique and experimental design. In section 4, we report findings from the behavioral experiments comparing human decision making accuracy with or without UQ information. In section 5, we report the impact of different visualizations of UQ information. Finally, sections 6 and 7 conclude by discussing the implications of our results and future work.

2. Background and Related Work

2.1. Human Decision-Making and Uncertainty

The possible benefit of AI UQ is supported by work in the judgment and decision-making literature on decision-

[1]DEVCOM Army Research Laboratory [2]University of Texas at Dallas, Richardson, TX. Correspondence to: Laura Marusich <laura.m.cooper20.civ@army.mil>, Yan Zhou <yan.zhou2@utdallas.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]We wish to highlight that for the classification task, ground truths for class labels are utilized to offer well-calibrated, high quality, instance-level uncertainty quantification for human subject experiments.

making under uncertainty. This work shows that providing overall prediction uncertainty enhances decision-making accuracy. For example, in weather forecasting, humans demonstrate higher decision-making performance when they receive well-calibrated probabilistic information (e.g., a forecast with a probability of rain), compared to only deterministic predictions (e.g., it will or will not rain) (Frick & Hegg, 2011; Joslyn & LeClerc, 2013; Morss et al., 2008; Nadav-Greenberg & Joslyn, 2009). However, increasing information, even when it is task-relevant, is not always beneficial to human decision-making performance (e.g., Marusich et al., 2016; Gigerenzer & Brighton, 2009; Alufaisan et al., 2021). An additional consideration is the way that uncertainty information is represented. In human decision-making, communicating uncertainty with visual representations and other intuitive methods can be especially effective (Gigerenzer et al., 2007; Hullman et al., 2018).

Despite previous general findings that uncertainty information is useful for decision-making, there is limited behavioral research assessing the benefits of AI UQ, particularly for human decision-making accuracy. Some existing qualitative work (e.g., Prabhudesai et al., 2023) suggests that the addition of UQ to predictions can impact the decision-making process of users and possibly reduce over-reliance on AI predictions. Among quantitative studies that do assess objective accuracy performance (e.g., Zhang et al., 2020; Buçinca et al., 2021), both the methods and results vary. In particular, the quality of the UQ calibration varies, with some studies opting to simulate AI prediction confidence with wizard-of-oz techniques, and others using the prediction probabilities generated by their model, but without quantifying the calibration of those probabilities. As a result, the potential benefits for AI UQ remain at least somewhat of an open question (Lai et al., 2021).

There is a clear gap for behavioral studies assessing human decision-making performance using quantifiably well-calibrated AI UQ for predictions. Our method for AI UQ uses known class labels to ensure high-quality uncertainty information at the instance-level, as poorly calibrated uncertainty information is likely to be detrimental to decision-making. We emphasize that *the application of known class labels to generate instance-level UQ aims to provide well-calibrated AI UQ for individual predictions specifically in the context of human subject experiments*. This approach is not *designed for real-life deployment scenarios where class labels may not be known in advance*. In the next section, we briefly provide context of existing techniques for AI UQ, which are often model-based and typically do not require labelled data.

2.2. Techniques for AI UQ

Predictions by AI-based systems are subject to uncertainty from different sources. The source of uncertainty is either aleatoric, caused by noise in data and irreducible, or epistemic because of uncertain model distribution (Kendall & Gal, 2017). Uncertainty quantification methods have been developed to assess the reliability of AI predictions (Abdar et al., 2021a), including Bayesian methods and ensemble methods (Abdar et al., 2021b).

Monte Carlo sampling (Neal, 2012) and Markov chain Monte Carlo (Salakhutdinov & Mnih, 2008; Salimans et al., 2015; Chen et al., 2014; Ding et al., 2014; Chen et al., 2015; Li et al., 2016; Gong et al., 2019) are heavily used for uncertainty quantification in Bayesian techniques (Kendall & Gal, 2017; Wang et al., 2019; Liu et al., 2019a). To estimate aleatoric uncertainty, a hidden variable is often proposed to represent the underlying data point x^* from which a given instance x is only one of many possible observations of x^*. Parameters modeling the transformation from x^* to x can be sampled to obtain multiple copies of the hidden x^*. For epistemic uncertainty, the distribution of model parameter θ is often approximated during training by achieving certain objective optimization, for example, the Kullback–Leibler divergence. The distribution of the prediction can be sampled from the samples of the learned model parameters. The predictive uncertainty can be established from the variance or entropy of the sampled predictions of the sampled hidden states of a given instance.

Quantifying uncertainty on learning models from a Bayesian perspective takes many different forms. Uncertainty Posterior distribution over Bayesian Neural Network (BNN) weights can be learned using variational inference (Subedar et al., 2019; Louizos & Welling, 2017; Farquhar et al., 2020; Ghosh et al., 2020). On the other hand, Generative Adversarial Networks (GANs) are used to generate out-of-distribution (OoD) examples (Oberdiek et al., 2022). Implicit neural representations (INRs) are reformulated from a Bayesian perspective to allow for uncertainty quantification (Vasconcelos et al., 2023). Similarly, Direct Epistemic Uncertainty Prediction (DEUP) is proposed to address the issue that using the variance of the Bayesian posterior does not capture the epistemic uncertainty induced by model misspecification (Lahlou et al., 2023). Aleatoric uncertainty and epistemic uncertainty have also been modeled as universal adversarial perturbations (Liu et al., 2019a).

Ensemble models can enhance the predictive accuracy, however, it is highly debated whether an ensemble of models can provide a good uncertainty estimate (Abdar et al., 2021b; Wilson & Izmailov, 2020; Sensoy et al., 2018). Recently, benefits of prior functions and bootstrapping in training ensembles with estimate of uncertainty have been discussed (Dwaracherla et al., 2023). Maximizing Overall

Diversity takes into account ensemble predictions for possible future input when estimating uncertainty (Jain et al., 2020). Random parameter initialization and data shuffling have also been proposed to estimate the uncertainty of DNN ensembles (Lakshminarayanan et al., 2017). A Bayesian non-parametric ensemble (BNE) approach is proposed to account for different sources of model uncertainty (Liu et al., 2019b). More details on extensive studies on quantifying uncertainty with respect to both Bayesian and ensemble methods, as well as in real applications can be found in (Abdar et al., 2021b). However, prediction probabilities are prone to overconfidence in some AI models. There is a lack of discussion on the calibration of uncertainty estimates in the existing literature.

Another related challenge is predictive multiplicity: models with similar performance yielding contradictory predictions (Watson-Daniels et al., 2023). One approach to resolving conflicting predictions is using their variations to calculate a risk score. Risk scores are typically point estimates, although there are exceptions such as the Viable Prediction Range over a set of models (Watson-Daniels et al., 2023). Here, we do not develop a novel UQ method. Instead our aim is assessing if well-calibrated UQ can improve human decision-making.

In this work, we achieved efficiency of UQ estimate by assessing the change of prediction yielded from repeatedly sampling noise adjacent to a given instance, and carefully calibrated the uncertainty information shown to the user by leveraging the ground truth. More precisely, *we provide well-calibrated uncertainty estimates in different visualizations of confidence intervals to the human participants*. Unlike the existing work discussed above, our goal is to *provide the uncertainty information to the human participants* to understand whether well-calibrated *uncertainty quantification information helps in user decision-making*. To achieve this goal, we do not attempt to come up with a UQ method a priori. Instead, we take the liberty of knowing the true labels of given instances, and simplify the problem as sampling predictive confidence from instances distorted with a small amount of random noise. The quality of the disclosed uncertainty estimate is verified using a strictly proper scoring rule (Gneiting & Raftery, 2007) prior to use in two behavioral experiments. While there have been recent calls for research using UQ with human decision-making (e.g., Bhatt et al. 2021; Lai et al. 2021), the few existing studies tend to focus on qualitative or subjective assessments of human behavior (e.g., Prabhudesai et al. 2023). Furthermore, it is not clear how useful to decision-makers the UQ information provided in these studies is, *due to lack of proper calibration*.

3. Current Work

We conducted two experiments to assess the effect of providing visualizations of AI prediction UQ information upon the accuracy and confidence of human decision-making. The first experiment compares performance when AI uncertainty is provided to performance when only an AI prediction, or no AI information at all, is provided. The second experiment compares decision-making performance for different representations of AI uncertainty. Our methods and results for the instance-level predictive UQ and behavioral experiments are fully reproducible. See the supplementary material for details and links.

In both experiments, we assessed our research questions using three different publicly-available and widely-used datasets: the *Census*, *German Credit*, and *Student Performance* datasets from the UCI Machine Learning Repository (Dua & Graff, 2017), described in more detail below.

3.1. Datasets

The *Census* dataset has 48,842 instances and 14 attributes. The missing values in the dataset were replaced with the mode (the most frequent value), and the dollar amounts were adjusted for inflation. The *German Credit* dataset has 1,000 instances and 20 attributes. The currency values were converted to dollars and adjusted for inflation. The *Student Performance* has 649 instances and 33 attributes. Three of the attributes *first period grade*, *second period grade*, and *final grade* were combined into one with their average. Each dataset was split into training (70%) and test (30%) data sets.

We selected these datasets because they involve real-world contexts that are fairly intuitive for non-expert human participants to reason about (e.g., will a student pass or fail a class?). In addition, using three datasets that vary in number of features and in the overall accuracy classifiers can achieve in their predictions ensures that our findings are not limited only to one specific dataset.

Several machine learning models were trained on all three datasets, including decision tree, logistic regression, random forest, and support vector machine. The best set of hyper-parameters was determined through grid search. Random forest was the best in terms of overall accuracy on the datasets and therefore was selected for use as the AI model in this study. The mean accuracy on the *Census* data is 85.3%, 75.7% on the *German Credit* data, and 85.1% on the *Student Performance* data. All classification tasks were completed on an Intel® Xeon® machine with a 2.30GHz CPU.

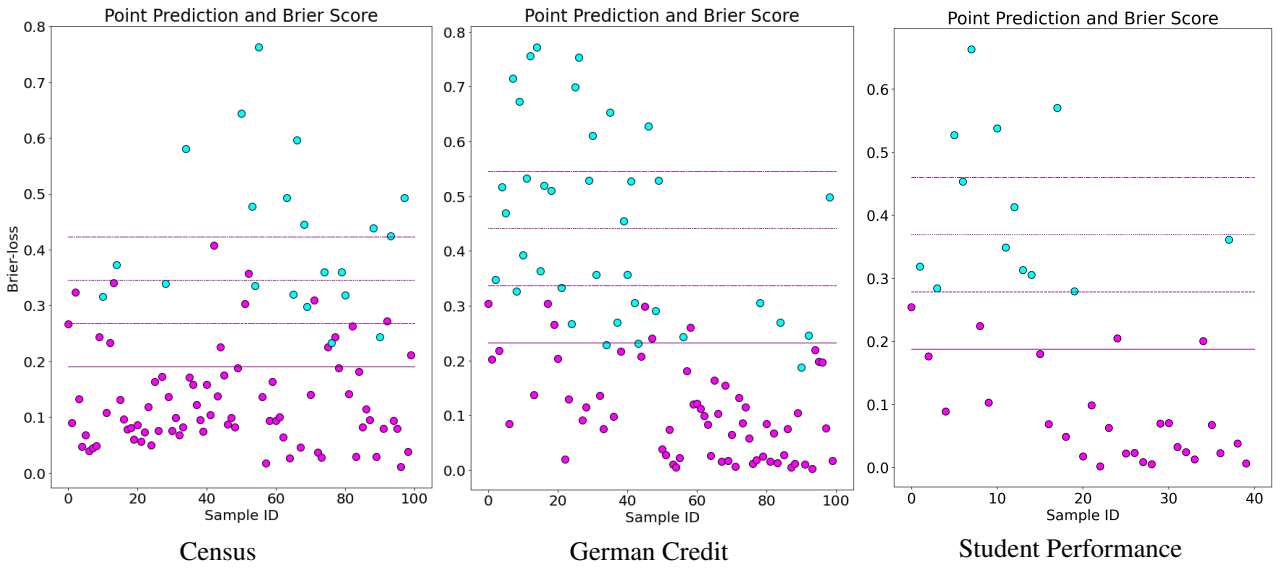

Figure 1. The Brier score of the "cloned" instances for *Census* (100), *German Credit* (100), and *Student Performance* (40) sampled for demonstration. Y-axis is the Brier score. Magenta marks the samples that are correctly predicted by the AI model, cyan marks samples incorrectly predicted by the model. Horizontal lines illustrate the mean of the Brier score, and its 0.5, 1, and 1.5 standard deviations.

3.2. Instance-Level Predictive Uncertainty Quantification

UQ methods in existing literature estimate predictive uncertainty without the knowledge of the true labels of the test instances. These methods are subject to complicated calculations, sometimes poor convergence, lack of scalability, and sometimes, they are time and resource consuming (Abdar et al., 2021a). In our study, we aim to provide predictive uncertainty quantification to human decision-makers and use the advantage of knowing the true labels in advance. Therefore, we simplify the problem as sampling predictive confidence from samples of x with a small random disturbance and *verify the quality of the uncertainty estimate* using a strictly proper scoring rule (Gneiting & Raftery, 2007) before showing it to the human. Note that, without knowing the ground truth, this treatment of UQ would be reckless and naive. It would appear that we model a prior distribution over hypothesis as the distribution over observations in the noisy neighborhood of a given instance. However, given the true label of an instance, we can hypothesize that observations over its n neighboring noisy samples are n plausible fits for this instance, and confirm our hypothesis with a strictly proper scoring rule.

Predictive uncertainty consists of data uncertainty (aleatoric) and model uncertainty (epistemic). To model data uncertainty, we sample n instances from a Gaussian distribution within a standard deviation σ from a given instance x, assuming $x = x^* + \eta$ where x^* is the clean input of x without the random disturbance η. Thus, given a prediction function parameterized by w, the class label of x is predicted as:

$$p(y|x, w) = \int p(y|x^*, w)p(x^*|x)dx^*$$

The posterior $p(x^*|x)$ is generally unknown. By assuming $\eta \sim \mathcal{N}(0, \sigma_0^2, I)$, we can sample from the posterior distribution given the noisy input x. In this study, we set $n = 100$ and $\sigma_0 = 0.1$.

Similarly, for model uncertainty, given a set of training data (X, Y), we assume there exists an uncertain set of m models with model uncertainty $\theta^{(m)} \sim p(\theta|X, Y)$. Hence, given an instance x, the probability of the class label of x is:

$$p(y|x, X, Y) = \mathbb{E}_{p(\theta|X,Y)}[p(y|x, \theta)].$$

In this study, we tested an ensemble of *logistic regression*, *support vector machine*, and *random forest* to predict the class label. The best uncertainty estimate, however, was obtained by using the *random forest* alone, assessed by the Brier score discussed below.

Predictive uncertainty per instance was computed for 294 randomly selected *Census* instances, 300 *German Credit* instances, and 194 *Student Performance* instances, for use in the behavioral study. Predictive uncertainty at the instance-level was measured on random samples in the neighborhood of the instance. More specifically, given an instance x, n random "clones" were sampled from a Gaussian distribution within δ standard deviation from the mean x. In the experiment, we let $n = 100$ and $\delta = 0.1$ which provided sufficient statistical significance and constrained neighborhood choices. Class probabilities were computed using the

trained random forest classifier for each of the 100 samples, and the 95% confidence interval of the class probabilities was used as the predictive uncertainty range for instance x. UQ computed from *random forest* alone was superior to that of the ensemble of *logistic regression*, *support vector machine*, and *random forest*, hence was used in the behavioral study.

Knowing the ground truth (class label) of the instances, we can verify the quality of the simulated predictive uncertainty using the Brier score (also referred to as Brier loss). The Brier score measures the mean squared difference between the predicted probability and the true outcome. For each selected instance x, with $y \in \{0, 1\}$ and the predicted probability $p_i = Pr(y_i = 1)$ for each "cloned" sample x_i, we compute the Brier score $B = \frac{1}{n}\sum_i (y - p_i)^2$ between the predicted probability of the "cloned" samples and y—the actual label of x.

If the "cloned" samples are truly representative of x, the computed Brier score should reflect the correctness of the prediction made for x by the AI model M. A smaller Brier score means more accurate predictions made for the clones of x, and therefore should correspond to a correct classification for x by M. We verified empirically that the Brier scores of the predictive uncertainty is highly correlated with the true prediction for x by M, as shown in Figure 1. Points with low Brier score corresponds to instances where M is correct. In Figure 1, points in magenta are the samples correctly predicted by the AI model, and points in cyan are samples incorrectly predicted by the model. As can be seen, "clones" for each correctly predicted sample correspond to low Brier score loss, and vice versa, cloned samples for incorrectly classified samples produce high Brier scores. Horizontal lines illustrate the mean of the Brier score, and its 0.5, 1, and 1.5 standard deviations. The Brier score close to mean (approximately 0.25) is a highly accurate indicator of the classification outcome. In essence, the Brier score resembles the trust score (Jiang et al., 2018) that has high precision at identifying correctly classified examples, and is adequate to assess the quality of the estimated UQ.

3.3. Behavioral Experiments: General Methods

We used the same experimental task across both Experiment 1 and 2, which was developed using jsPsych (De Leeuw, 2015) and hosted on MindProbe https://mindprobe.eu/ using Just Another Tool for Online Studies (JATOS) https://github.com/JATOS/JATOS. Each trial of this task included a description of an individual and a two-alternative forced choice for the classification of that individual. Each choice was correct on 50% of the trials, thus chance performance for human decision-making accuracy was 50%. In some conditions, an AI prediction or an AI prediction and a visualization of prediction uncertainty would also appear. Figure 2 shows an example of the information appearing in the three AI conditions for a trial from the German Credit dataset condition (see supplementary material for more example trials). After making a decision, participants then entered their confidence in that choice, on a Likert scale of 1 (No Confidence) to 5 (Full Confidence). Feedback was then displayed, indicating whether or not the previous choice was correct.

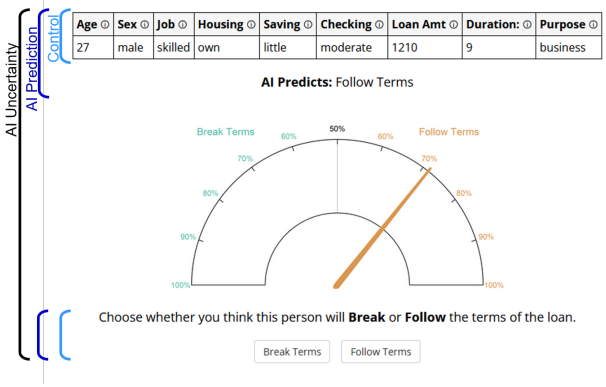

Figure 2. Example showing the information appearing in the three AI conditions in Experiment 1 for a trial from the *German Credit* dataset condition.

For each dataset, we selected 50 instances with representative average AI prediction accuracies (*Census*: 88%, *German Credit*: 76%, *Student Performance*: 82%). Then, for each participant, we randomly sampled 40 of those 50 instances for the block of test trials, resulting in small variations in AI accuracy for each participant.

Online participants were recruited from Prolific (https://www.prolific.co). They provided informed consent, viewed a series of instructional screens with examples, completed 8 practice trials, followed by 40 test trials, and a brief series of questionnaires (demographics, self-reported strategies, subjective usability, subjective task difficulty, task understanding, and an assessment of risk literacy (Cokely et al., 2012), see supplementary material). Most participants completed the task in less than 20 minutes, and they were paid $5.00 for their participation (i.e., well above the U.S. federal minimum hourly wage). This research received Institutional Review Board (IRB) approval.

4. Experiment 1

Experiment 1 compared participant decision-making accuracy in three conditions: Control (no AI prediction information), AI Prediction, and AI Uncertainty (AI prediction plus a visualized point estimate of AI uncertainty), see Figure 2. All hypotheses and methods were pre-registered (https://aspredicted.org/ZW9_Z54).

We hypothesized that for all three datasets, participant decision accuracy would be highest in the AI Uncertainty condition, followed by the AI Prediction condition, and lowest in the Control condition. Similarly, we hypothesized that confidence calibration (positive association between confidence and accuracy) would be strongest in the AI Uncertainty condition, followed by the AI Prediction condition, and lowest in the Control condition.

4.1. Participants

We recruited nearly 50 participants in each of 9 experimental conditions, for a total of 445 participants (48.8% male, 48.5% female, 2.7% other or prefer not to answer). The majority (68.8%) of participants were 18-44 years old.

4.2. Results and Discussion

We excluded trials with reaction times that exceeded three standard deviations above the mean; this resulted in the removal of 396 out of 17,800 trials across all participants. An omnibus 3 (AI Condition) x 3 (Dataset) ANOVA for mean accuracy indicated a significant main effect of AI condition ($F(2, 436) = 84.11, p < 0.001, \eta_p^2 = 0.28$; see left side of Figure 3). This large effect size is driven primarily by the differences between the Control condition and the other two conditions. However, we also used Tukey's honest significance test to conduct post-hoc comparisons between individual conditions. These comparisons showed not only that accuracy in the AI Prediction condition was higher than in the Control condition ($t(436) = 9.91, p < 0.0001$), but also that accuracy in the AI Uncertainty condition was further improved (although to a lesser extent) over the AI Prediction condition ($t(436) = 2.36, p = 0.049$).

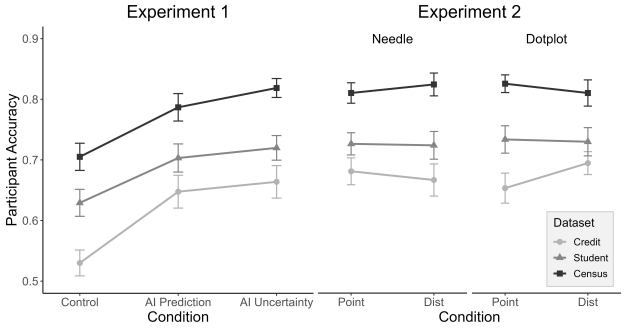

Figure 3. Participant accuracy in Experiments 1 (left) and 2 (right). Error bars represent 95% confidence intervals.

There was also a significant main effect of dataset upon accuracy ($F(2, 436) = 144.72, p < 0.001, \eta_p^2 = 0.40$). Unsurprisingly, human accuracy was lowest in the *German Credit* dataset, for which AI accuracy is also relatively low,

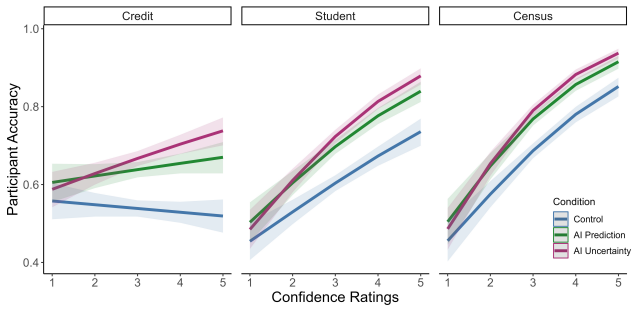

Figure 4. Predicted effects (level 2/overall results of multilevel model) of confidence ratings, dataset, and AI condition upon accuracy in Experiment 1. Steeper, positively-sloped lines indicate better confidence calibration. Shaded areas represent 95% confidence intervals for the predicted values.

and human accuracy was highest in the *Census* dataset, where AI accuracy is also relatively high.

To assess confidence calibration, we fit a multilevel model (Gelman & Hill, 2006) that included dataset, AI condition, and confidence ratings as fixed-effect predictors, with varying intercepts for each participant. The multilevel model accounts for a moderate amount of variance in fixed and varying effects, conditional Pseudo-R^2 = 0.10. This model indicated that the relationship between confidence and accuracy interacted with both dataset and AI condition, illustrated in Figure 4.

As hypothesized, confidence was most highly calibrated with accuracy in the AI Uncertainty condition, followed by the AI Prediction condition, and lowest in the Control condition. See data and code links in the supplementary material for details of the model.

We also analyzed the impact of AI condition and dataset upon participants' response times (RT). Using an omnibus 3 x 3 ANOVA, we found a significant main effect of dataset upon RT ($F(2, 436) = 6.22, p = 0.002, \eta_p^2 = 0.03$), with participants responding slowest in the *Student Performance* dataset (see left side of Figure 5), perhaps due to the larger number of attributes to consider for each instance. We did not find significant effects of AI condition ($F(2, 436) = 1.97, p = 0.14, \eta_p^2 = 0.009$) or an interaction effect ($F(4, 436) = 0.13, p = 0.97, \eta_p^2 = 0.001$). These results imply that the accuracy benefit for AI UQ information is not merely due to a speed/accuracy tradeoff among participants (Wickelgren, 1977).

5. Experiment 2

Experiment 1 demonstrated that decision-making performance can be improved with AI UQ information; we de-

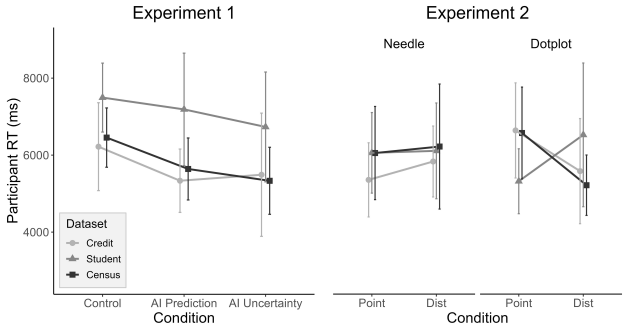

Figure 5. Participant response times (RT) in milliseconds in Experiment 1 (left) and Experiment 2 (right). Error bars represent 95% confidence intervals.

signed Experiment 2 to test if different representations of UQ might be more or less beneficial for decision-making. We compared performance with distributions of uncertainty probabilities to point-estimated probabilities, as well as two different visualizations of uncertainty (needle vs. dotplot), again using the same three datasets used in Experiment 1 (see Figure 6). All hypotheses and methods were pre-registered (https://aspredicted.org/CJW_71H).

We hypothesized that for all three datasets, both participant decision accuracy and confidence calibration would be higher with distribution information than for point-estimated UQ in the AI Uncertainty condition, followed by the AI Prediction condition, and lowest in the Control condition. We also hypothesized that, within the distribution conditions, accuracy and confidence calibration would be higher for the dotplot visualization than for the needle visualization, due to the more detailed information about the shape of distributions available in the dotplot, compared to the needle which only shows the distributions as a uniform range.

5.1. Participants

We recruited 50 participants in each cell, for a total of 600 participants (48.5% male, 49.3% female, 2.2% other or prefer not to answer), from the Prolific platform. Most participants (75.3%) were 18-44 years old.

5.2. Results and Discussion

We excluded 553 trials (out of 24,000 across all participants) with reaction times that exceeded three standard deviations above the mean. An omnibus 2 (point vs. distribution) x 2 (needle vs. dotplot) x 3 (dataset) ANOVA for mean accuracy indicated a significant main effect of dataset ($F(2, 588) = 188.77, p < 0.001, \eta_p^2 = 0.39$), where performance was again highest for the *Census* dataset, followed by *Student Performance*, and lowest for *German Credit*. However, there was no significant main effects of point vs. distribution ($F(1, 588) = 0.28, p = 0.60, \eta_p^2 < 0.001$) or visualization type ($F(1, 588) = 0.16, p = 0.69, \eta_p^2 < 0.001$). Neither was there evidence of significant first-order interaction effects among the three manipulated variables, see right side of Figure 3).

As in Experiment 1, we fit multilevel models with varying intercepts for each participant to assess confidence calibration. We found that the best-fitting model (as assessed by fit statistics: AIC and BIC) had only confidence, dataset, and their interaction as fixed-effect predictors, see Figure 7.

The best fit model indicates confidence calibration, again, accounts for a moderate amount of variance in fixed and varying effects, conditional Pseudo-R^2 = 0.13. Including point vs. distribution or visualization type did not appear to improve model fit or to predict accuracy performance above and beyond what is predicted by dataset and confidence. See data and code links in the supplementary material for details of the model.

Thus, contrary to the pre-registration, neither our accuracy or confidence calibration results indicate support for our Experiment 2 hypotheses. The hypotheses were that performance would be better with distribution information for UQ than for point-estimated UQ, and that within the distribution condition, performance would be better with dotplots than with the needle visualization.

Additionally, we analyzed RT in Experiment 2 (see right side of Figure 5) to rule out speed/accuracy tradeoffs in the performance results. An omnibus 2 x 2 x 3 ANOVA indicated no significant effect of dataset ($F(2, 588) = 0.09, p = 0.92, \eta_p^2 < 0.001$), point vs. distribution ($F(1, 588) = 0.06, p = 0.81, \eta_p^2 < 0.001$), or visualization type ($F(1, 588) = 0.01, p = 0.91, \eta_p^2 < 0.001$) upon RT. Neither was there evidence of significant interaction effects among the three manipulated variables.

5.3. Exploratory Analyses

Although this work was not specifically designed to assess whether the combination of humans plus AI generally exceeded the accuracy of AI predictions, we conducted exploratory analyses to investigate how often this occurred. In both experiments, we found that a small subset of participants were able to outperform the accuracy of the AI predictions they received (see Table 1). AI uncertainty enabled humans to outperform the AI accuracy more frequently and with patterns of mean differences suggesting greater improvements.

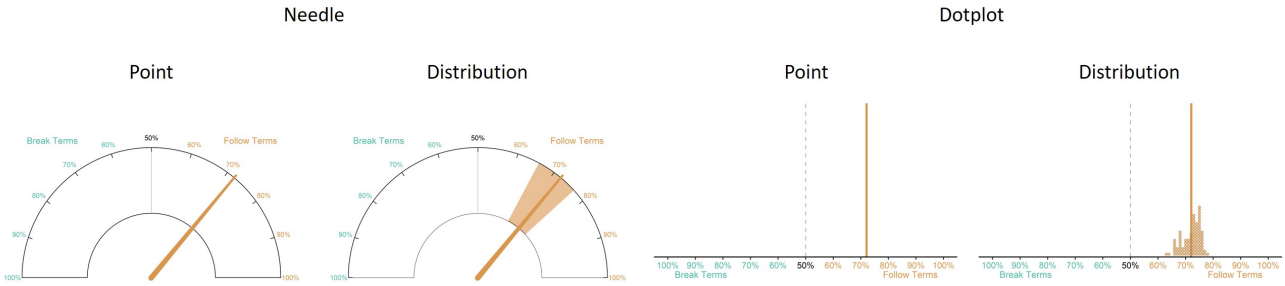

Figure 6. Example of the four conditions (point vs. distribution, and needle vs. dotplot) in Experiment 2, using a trial from the *German Credit* dataset.

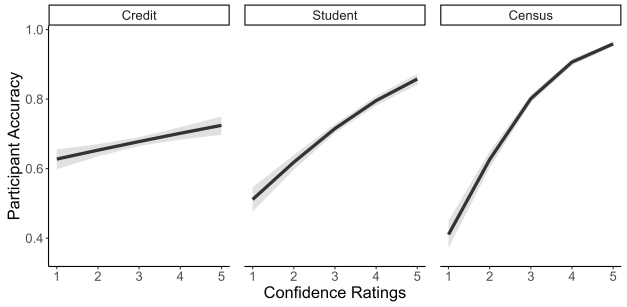

Figure 7. Predicted effects (level 2/overall results of multilevel model) of confidence ratings and dataset upon accuracy in Experiment 2. Steeper, positively-sloped lines indicate better confidence calibration. Shaded areas represent 95% confidence intervals for the predicted values.

Table 1. Number of participants who outperformed the AI

Human Accuracy	>AI (n)	≤AI (n)	>AI (%)
Experiment 1			
AI Prediction	2	148	1.33%
AI Uncertainty	7	141	4.73%
Experiment 2			
Point Needle	8	142	5.33%
Point Dotplot	8	142	5.33%
Dist Needle	11	139	7.33%
Dist Dotplot	7	143	4.67%

6. General Discussion, Limitations, and Future Work

The overall results of Experiment 1 showed that providing AI UQ to human decision-makers improved accuracy and confidence calibration performance over and above providing an AI prediction alone. In Experiment 2, we did not find meaningful differences between different representations of UQ. Taken together, our findings suggest that the *benefit of AI UQ may not be overly sensitive to the representation of the UQ information*. Also, adding more information (here, the UQ distribution), even though task-relevant, did not improve decision-making which is consistent with prior research (Marusich et al., 2016; Gigerenzer & Brighton, 2009; Alufaisan et al., 2021).

Here, the humans and AI had identical information so we could evaluate well-calibrated UQ. Normatively, algorithms tend to outperform human decision-making in a variety of tasks; a clear exception is when people have knowledge the algorithm does not (Dawes et al., 1989). In such situations, it is possible that the AI and human combined can produce better performance than either alone (Cummings, 2014). Exploratory results suggested AI uncertainty may increase the frequency of people exceeding AI accuracy, although this was not a common occurrence; likely because both the human and the AI had the same information.

In this work, we did not compare different UQ techniques but did demonstrate that, using a Brier score, our UQ technique performed well in practice. It is possible that UQ techniques that do not perform as well as ours may not improve human decision-making accuracy. In future work, we plan to use other UQ techniques, including less effective ones to understand the impact of UQ quality on human decision-making accuracy. Similarly, behavioral experiments were limited to comparing point versus distributions and two visualizations of predictive uncertainty. It is possible that other visualizations may make providing distribution information more effective, and we plan to conduct future work assessing more UQ visualization techniques.

Finally, we used a relatively simple binary classification decision-making task. In more complex application domains, where humans encounter multi-class classification problems, the impact of UQ information on human decision making could be different. Future work should explore multi-class classification problems, as well as the use of AI

UQ for complex tasks where people, such as experts, have knowledge unknown to the AI model.

7. Conclusion

Our extensive behavioral experiments show that providing *high quality* AI uncertainty information improves human decision accuracy and confidence calibration over the AI prediction alone. This human performance benefit was not limited to only a specific visual representation of UQ information. In previous work, there is an absence of evaluating calibration for AI UQ (Lai et al., 2021). Here, we used Brier scores to quantify the calibration of our implementation of AI UQ. We showed the AI UQ used here was well-calibrated by leveraging the existing class labels of test instances.

Software and Data

See supplementary material at https://osf.io/cb762/.

Impact Statement

Recently, understanding how humans and AI systems can work better together has emerged as an important challenge. Although previous work has explored how and when explainable AI may help human decision making, the impact of providing uncertainty information to humans has not been explored in depth in the context of AI systems. To address this challenge, in this work, we explore whether providing uncertainty information to humans may improve decision making and potentially correct errors caused by the AI models.

Acknowledgements

We thank Mary Grace Kozuch for helpful feedback on a previous version of this paper.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the U.S. DEVCOM Army Research Laboratory (ARL) or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation. M.K. and Y.Z. were supported in part by NSF awards DMS-2204795, OAC-2115094, CNS-2331424, ARL/Army Research Office award W911NF-17-1-0356, NIH award 5RM1HG009034-08, National Center for Transportation Cybersecurity and Resiliency (TraCR) award and a gift from Cisco Inc.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76 (C):243–297, dec 2021a. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.008. URL https://doi.org/10.1016/j.inffus.2021.05.008.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021b.

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., and Kantarcioglu, M. Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6618–6626, May 2021. doi: 10.1609/aaai.v35i8.16819. URL https://ojs.aaai.org/index.php/AAAI/article/view/16819.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–13, 2019.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., and Xiang, A. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 401–413, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462571. URL https://doi.org/10.1145/3461702.3462571.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021. doi: 10.1145/3449287. URL https://doi-org.ezproxy.uta.edu/10.1145/3449287.

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2278–2286, Cambridge, MA, USA, 2015. MIT Press.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient hamiltonian monte carlo. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1683–1691, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/cheni14.html.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., and Garcia-Retamero, R. Measuring risk literacy: The berlin numeracy test. *Judgment and Decision making*, 7(1): 25–47, 2012.

Cummings, M. M. Man versus machine or man+ machine? *IEEE Intelligent Systems*, 29(5):62–69, 2014.

Dawes, R. M., Faust, D., and Meehl, P. E. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989.

De Leeuw, J. R. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47:1–12, 2015.

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D., and Neven, H. Bayesian sampling using stochastic gradient thermostats. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/21fe5b8ba755eeaece7a450849876228-Paper.pdf.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dwaracherla, V., Wen, Z., Osband, I., Lu, X., Asghari, S. M., and Roy, B. V. Ensembles for uncertainty estimation: Benefits of prior functions and bootstrapping. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=IqJsyulDUX.

Farquhar, S., Smith, L., and Gal, Y. Try depth instead of weight correlations: Mean-field is a less restrictive assumption for deeper networks. *CoRR*, abs/2002.03704, 2020. URL https://arxiv.org/abs/2002.03704.

Frick, J. and Hegg, C. Can end-users' flood management decision making be improved by information about forecast uncertainty? *Atmospheric Research*, 100(2-3):296–303, 2011.

Gelman, A. and Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

Ghosh, P., Sajjadi, M. S. M., Vergari, A., Black, M. J., and Schölkopf, B. From variational to deterministic autoencoders. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1g7tpEYDS.

Gigerenzer, G. and Brighton, H. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, 1(1):107–143, 2009.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 8(2):53–96, 2007.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Gong, W., Tschiatschek, S., Nowozin, S., Turner, R. E., Hernández-Lobato, J. M., and Zhang, C. Icebreaker: Element-wise efficient information acquisition with a bayesian deep latent gaussian model. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c055dcc749c2632fd4dd806301f05ba6-Paper.pdf.

Hullman, J., Qiao, X., Correll, M., Kale, A., and Kay, M. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics*, 25(1):903–913, 2018.

Jain, S., Liu, G., Mueller, J., and Gifford, D. Maximizing overall diversity for improved uncertainty estimates in deep ensembles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 4264–4271, Apr. 2020. doi: 10.1609/aaai.v34i04.5849. URL https://ojs.aaai.org/index.php/AAAI/article/view/5849.

Jalaian, B., Lee, M., and Russell, S. Uncertain context: Uncertainty quantification in machine learning. *AI Magazine*, 40(4):40–49, 2019.

Jiang, H., Kim, B., Guan, M. Y., and Gupta, M. To trust or not to trust a classifier. In *Proceedings of the 32nd International Conference on Neural Information Processing*

Systems, NIPS'18, pp. 5546–5557, Red Hook, NY, USA, 2018. Curran Associates Inc.

Joslyn, S. and LeClerc, J. Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, 22(4):308–315, 2013.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

Lahlou, S., Jain, M., Nekoei, H., Butoi, V. I., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=eGLdVRvvfQ.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Li, C., Stevens, A., Chen, C., Pu, Y., Gan, Z., and Carin, L. Learning weight uncertainty with stochastic gradient mcmc for shape classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5666–5675. IEEE Computer Societyhelp@computer.org, December 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.611. Generated from Scopus record by KAUST IRTS on 2021-02-09.

Liu, H., Ji, R., Li, J., Zhang, B., Gao, Y., Wu, Y., and Huang, F. Universal adversarial perturbation via prior driven uncertainty approximation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 2941–2949. IEEE, 2019a. doi: 10.1109/ICCV.2019.00303. URL https://doi.org/10.1109/ICCV.2019.00303.

Liu, J., Paisley, J., Kioumourtzoglou, M.-A., and Coull, B. Accurate uncertainty estimation and decomposition in ensemble learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1cc8a8ea51cd0adddf5dab504a285915-Paper.pdf.

Louizos, C. and Welling, M. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 2218–2227. JMLR.org, 2017.

Marusich, L. R., Bakdash, J. Z., Onal, E., Yu, M. S., Schaffer, J., O'Donovan, J., Höllerer, T., Buchler, N., and Gonzalez, C. Effects of information availability on command-and-control decision making: Performance, trust, and situation awareness. *Human Factors*, 58(2):301–321, 2016. doi: 10.1177/0018720815619515. URL https://doi.org/10.1177/0018720815619515. PMID: 26822796.

Morss, R. E., Demuth, J. L., and Lazo, J. K. Communicating uncertainty in weather forecasts: A survey of the us public. *Weather and forecasting*, 23(5):974–991, 2008.

Nadav-Greenberg, L. and Joslyn, S. L. Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, 3(3):209–227, 2009.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Oberdiek, P., Fink, G. A., and Rottmann, M. UQGAN: A unified model for uncertainty quantification of deep classifiers trained via conditional GANs. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=djOANbV2zSu.

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q. V., and Banovic, N. Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-ai decision making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pp. 379–396, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701061. doi: 10.1145/3581641.3584033. URL https://doi.org/10.1145/3581641.3584033.

Salakhutdinov, R. and Mnih, A. Bayesian probabilistic matrix factorization using markov chain monte carlo.

In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 880–887, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390267. URL https://doi.org/10.1145/1390156.1390267.

Salimans, T., Kingma, D., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1218–1226, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/salimans15.html.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., and Vössing, M. A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 617–626, 2022.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 3183–3193, Red Hook, NY, USA, 2018. Curran Associates Inc.

Subedar, M., Krishnan, R., Meyer, P. L., Tickoo, O., and Huang, J. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Vasconcelos, F., He, B., Singh, N. M., and Teh, Y. W. UncertaINR: Uncertainty quantification of end-to-end implicit neural representations for computed tomography. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=jdGMBgYvfX.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *NEUROCOMPUTING*, 338:34–45, April 2019. ISSN 0925-2312. doi: 10.1016/j.neucom.2019.01.103.

Watson-Daniels, J., Parkes, D. C., and Ustun, B. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10306–10314, 2023.

Wickelgren, W. A. Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41(1):67–85, 1977.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 295–305, 2020.